Lecture Notes, Set 4

This set of notes covers hypothesis testing with an emphasis on p-values. Together with confidence intervals, this represents by far the most important of the subjects we are reviewing from Math 218.

Introduction To Hypothesis Tests

The subject of the next section of notes is hypothesis testing. This is usually one of the hardest topics for students to grasp, so I will try to give you a number of different ways to view hypothesis testing problems. The basic idea is this. You have some theory that tells you what the value of a population mean (or other parameter) should be. You want to decide whether that theory is right or not. You make a decision based on available sample data. The following are examples of hypothesis testing situations:

Example 1: I have a coin which I claim is fair. I toss the coin 100 times and report to you the number of heads I see. Based on the observed number of coins, you decide whether or not my coin is really fair. If the coin is truly fair, you expect to see around 50 heads. That is about half the tosses should come up heads. If I told you that I had gotten 51 heads, or 48, it probably wouldn't make you suspicious about the fairness of my coin. However, if I got 100 heads in a row, you definitely wouldn't believe the coin was fair. Why not? Because, IF THE COIN IS FAIR, I AM EXTREMELY UNLIKELY TO GET 100 HEADS, whereas 48 heads or 51 heads is a perfectly reasonable number. The difference from 50 is probably just random fluctuation. This is the basic idea of hypothesis testing. Start with a theory–e.g. the coin is fair. Look at some data. If the data are consistent with the theory, you accept the theory. If the data are very unlikely ASSUMING the theory is true, you reject the theory.

Example 2: Management wants to improve the speed with which workers on an assembly line perform a certain task. They take a random sample of n=25 workers, and record how long it takes the workers to perform the task. Then they administer a training program and measure the time to complete the task again. Let X_i be the reduction in time to complete the task for the ith worker. That is, $X_i = 10$ means the worker completed the task 10 minutes faster. For the sample of 25 workers, the average reduction was $\bar{X} = 2.5$ minutes, and the standard deviation was $s = 9$ minutes. Management would like to know if the training program helped or hurt the workers. This is also a hypothesis testing situation. One theory is that the workers did better after the training program. The alternative theory is that the training made no difference, or even slowed the workers down. A high positive value of $\bar{X}$ would convince you the training helped, whereas a small or negative value would be evidence that the training was no good. Note that this situation is a bit different from Example 1. There either a very high or very low value discredited the original theory. Here only a very high value convinces you that the training helps.

There are many types of hypothesis testing situations. Phrases which should tip you off include "test the hypothesis at level α....", "is the data consistent with....", "does treatment a work better than treatment b", "is group a different from group b..." etc.

Two Ways To View Hypothesis Tests: Confidence Intervals and Probabilities

We performed hypothesis tests earlier without even being aware that we were doing it. One way to perform an hypothesis test is to compute a confidence interval. Remember that a confidence interval for the population mean, gives you a range of "plausible" values for the true mean μ. Sometimes there is a particular

value of mu (or of p, the population proportion) which is of special interest. Here are some examples:

Example 1: In an election, a candidate wins if they get more than 50% of the vote. Suppose a poll was taken before an election to see what proportion of people supported Candidate A. The value of special interest to Candidate A is p=.5. Specifically Candidate A wants to know whether $p \leq .5$–i.e. the race is a dead heat, or he is losing–or whether $p \geq .5$–i.e. he is going to win.

Example 2: A company produces computer chips. A certain percentage, p, of their chips are defective. If more than 5% of the chips are defective customers will complain and the company will start losing money. The special value of p is $p = .05$. Specifically, the company wants to know if $p \leq .05$–i.e. everything is fine–or $p \geq .05$, in which case they need to fix what is wrong with the production process.

Example 3: The state of California has mandated that average smog levels in Los Angeles be under 100 units by the year 2000. The city measures smog levels at various points and times to see if they are in compliance with the mandate. The special value of μ, the true mean smog level is $\mu = 100$. If $\mu \leq 100$ the city is in compliance. Otherwise, if $\mu > 100$ the city will need to implement new smog reduction techniques.

Example 4: The FDA requires that canned food contain fewer than 5 micrograms of toxic substances. To see whether a company is in compliance with the regulations, the FDA tests 100 cans for toxic materials. They want to determine whether $\mu \leq 5$–i.e. the company is in compliance or whether $\mu > 5$ in which case the company will be fined and the product removed from the shelves.

Example 5: In the past, and average household has purchased 5.5 quarts of laundry detergent per year. A government board which monitors consumption of various products wants to know if the amount of laundry detergent used by Americans has changed in the last 20 years. The special value of μ in this case is 5.5. The board wishes to test whether $\mu = 5.5$–that is consumption levels have remained unchanged–or whether $\mu \neq 5.5$–i.e. consumption levels have changed.

How do we go about answering questions like these? There are two approaches. The first is to use confidence intervals. Given sample data, you can calculate a confidence interval for the true population mean or proportion. You can then see if the value of special interest to you lies in the interval. Here is an example:

Example: We had an example in which n=100 voters were surveyed before an election. Of those 100 voters, 64 supported Candidate A. We computed a 95% confidence interval for p, the true proportion of voters supporting Candidate A. Using the fact that $z_{\alpha/2} = z_{.025} = 1.96$ and that $\hat{p} = .64$ our confidence interval was

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$$

or

$$.64 \pm (1.96)(\sqrt{\frac{(.64)(.36)}{100}})$$

The resulting confidence interval is [.546, .754]. In other words, we are 95% sure that between 54.6 and 75.4% of voters support Candidate A. This means $p > .5$ and we are sure Candidate A is in the lead. We have just tested $p \leq .5$ versus $p > .5$!

Another way to do hypothesis tests is via a probability calculation. Consider the following problem:

Example: A company fires 40% of their work force, supposedly without regard to race. The minority workers at the company are suspicious because out of the n=200 minority workers, 100 (or 50%) were fired. The minority workers want to test whether the probability of a randomly selected individual worker being fired was really p=.4 as the company claimed. They compute the probability that if the true proportion

were p=.4, as many as 100 out of 200 minority workers would have been fired. The probability is

$$P(\hat{p} \geq .5) = P(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{.5 - p}{\sqrt{p(1-p)/n}}) = P(Z \geq \frac{.5 - .4}{\sqrt{(.4)(.6)/200}}) = P(Z \geq 2.89) = .0019$$

It is VERY unlikely that 100 out of 200 minority workers would have been fired if the workers each had a p=.4 chance of being fired. Therefore, the company probably discriminated against the minority workers. We have tested the hypothesis $p \leq .4$ versus the alternative $p > .4$. We did a similar problem earlier involving minorities waiting at a border crossing and reached the same type of conclusion.

The Intuition Behind Hypothesis Tests

The following represents an intuitive approach to hypothesis testing problems. The first step is to identify what value of μ or p is of critical interest to you. For instance, in the above problem, the critical value is p=.4. In the election problem, the critical value was p=.5. The next step is to identify the two alternatives. For instance in the coin example, the alternatives are that the coin is fair or that it isn't. In the assembly line example, training either helps or hinders. Generally, one of these alternatives is more interesting or important than the other. The first alternative is called the **null hypothesis** and is denoted H_0. There are several ways to characterize the null hypothesis. It can be

- The thing you want to disprove
- The status quo–that is that nothing has changed from the past
- Your default position–i.e. the thing you would assume unless someone provided strong evidence to the contrary
- The less interesting or important situation, or the one that does not require taking any action

The other hypothesis is called the **alternative hypothesis** and is denoted H_A. There are also several ways to characterize the alternative. It can be

- The thing you want to prove
- That what was true in the past is no longer true
- That something interesting or important or requiring action has occurred

Here are the null and alternative hypotheses for the examples we considered earlier:

Example: The election. The candidate wants to establish that he or she is ahead. Therefore the null hypothesis is that he or she is behind, and the alternative is that he or she is ahead. In symbols, we have $H_0 : p \leq .5$ versus $H_A : p > .5$.

Example: Defective products. If the machinery is running properly, nothing needs to be done. However, if more than 5% defective products are being produced it is critically important to find this out and stop it. Therefore, the null hypothesis is that the machines are working fine, and the alternative is that they are not. In symbols, we have $H_0 : p \leq .05$ versus $H_A : p > .05$. Note: if the company was trying to prove to an inspector that its process was working properly then the null and alternative hypotheses would be reversed. What you need to establish plays a big role in determining which is the null and which is the alternative.

Example: Smog in LA. The city wants to prove that it is meeting the standards. Therefore, the null hypothesis is that they are not meeting the standards and the alternative is that they are meeting the standards. In symbols, we have $H_0 : \mu \geq 100$ versus $H_A : \mu \leq 100$. If a government inspector was trying to

prove that the city was in violation of the policy, the roles of the two hypotheses would be switched.

Example: FDA. The company wants to prove that it is meeting the standards. Therefore, the null hypothesis is that they are not meeting the standards and the alternative is that they are meeting the standards. In symbols, we have $H_0 : \mu \geq 5$ versus $H_A : \mu \leq 5$. If a government inspector was trying to prove that the company was in violation of the policy, the roles of the two hypotheses would be switched.

Example: Detergent Consumption. The null hypothesis is that the status quo is being maintained–people still buy an average of 5.5 quarts of laundry detergent. The alternative is that detergent consumption has changed. If consumption has changed, companies that sell detergent may need to change their marketing practices. In symbols, $H_0 : \mu = 5.5$ versus $H_A : \mu \neq 5.5$.

Note that you start by **assuming the null hypothesis, the less interesting or important hypothesis, is true!** This may seem backwards if what you want to prove is that the alternative hypothesis is right. The reason for doing this is that it is much easier to prove something wrong than to prove something right. A single counter-example can prove a statement is false, but no single example will prove a statement is true.

The next step is to try to understand what it means for the problem at hand if the null hypothesis is true. For instance, in the coin example, the null hypothesis is that the coin is fair. If this is true then we should expect to see heads about half the time, or around 50 heads in 100 tosses. In the assembly line example, the null hypothesis is that the training program makes no difference (or even slows the workers down). If this is true we expect the average reduction in work time to be about zero (or even negative). The next step is to decide whether the data you actually observed is likely if the null hypothesis is true. For instance, in the coin example if I actually observed 52 heads in 100 tosses, this would seem perfectly consistent with the null hypothesis that the coin is fair. However, if the coin is really fair, then I am very unlikely to get just 2 heads. Thus an observed number of 2 heads would make me doubt the null hypothesis. Specifically, you calculate the probability of what you observed (or something even more extreme–that is even more favoring the alternative hypothesis) under the assumption that the null hypothesis is true. This probability is called a **p-value.** In the discrimination example above, the probability we calculated was a p-value.

If the value you observed is very unlikely given the null hypothesis, you **reject the null hypothesis.** If the value you saw is reasonable given the null hypothesis, you **fail to reject the null hypothesis.** Finally, you interpret your decision in the framework of the problem. For instance, in the case of the assembly line training program, you would decide whether or not it was worth implementing the training for all employees.

How do you decide whether or not the observed data is consistent with the null hypothesis? The usual way is to compute a probability. Suppose I am looking at the coin example. I observe 80 heads. I can compute the probability that I see a number of heads as extreme as 80, or even more extreme, when the coin is fair. What is "as extreme as 80?" It means as far away from the hypothesized number of heads (50) as 80. Thus values of 0, 1,...,20 ,80, 81,..., 100 are all as extreme or more extreme than 80. They all make me doubt the coin is fair as much or more as a value of 80 would do. I can calculate the probability of getting 20 or fewer or 80 or more heads using the binomial formula. This probability is virtually 0! Therefore, I am very unlikely to have seen 80 heads (or a more extreme value) if the coin were truly fair, and so I reject the null hypothesis. How small a probability is small enough to reject the null hypothesis? The rule of thumb is that 5% or less is small enough. However, there is nothing sacred about 5% and the actual value you choose should depend on context.

Formal Hypothesis Tests

What we have done so far represents an informal approach to hypothesis testing. I will now write down the formal procedure. However, I think it will be very helpful to keep the intuitive ideas given above in mind. The procedure I outline here depends only on the idea of p-values. Those of you who took Math 218 will

probably also remember using "rejection regions." I will not use rejection regions for several reasons. First, I think they are confusing and add an extra step to the calculation process. Second, in the future we will be doing all our hypothesis tests on the computer and the computer gives p-values, not rejection regions.

- **Step I** Identify the value of μ or p that is of special interest for you. For instance in the election problem the value p=.5 is of special interest because it corresponds to being ahead or behind. In the smog problem the value $\mu = 100$ was critical because it was the cutoff between compliance and non-compliance.

- **Step II: Establish the null hypothesis, H_0 and the alternative hypothesis, H_A.** This must be done both at an intuitive level and a mathematical level. For example, in the problem where I toss a coin 100 times and counted the number of heads, my null hypothesis, in English, is that the coin is fair, and my alternative hypothesis is that it is not fair. In mathematical symbols, the null hypothesis is $H_0 : p = .5$ and the alternative hypothesis is $H_A : p \neq .5$.

- **Step III (Optional) : Decide on the significance level, alpha for the test.** There are two kinds of errors you can make when you perform an hypothesis test. One is to reject the null hypothesis when the null hypothesis is correct. This is called **Type I error.** The amount of Type I error you are willing to allow is labeled α and is referred to as the **significance level** of the test. The second type of mistake you can make is to fail to reject the null hypothesis when in fact the null hypothesis is false. This is called **Type II error** and is denoted β. The number $1-\beta$ is called the **power** of the test because it tells you how often you are able to detect that the alternative hypothesis is correct when it is. Obviously you want the probabilities of Type I and Type II errors to be very small. Unfortunately, there is a tradeoff. You can make the probability of Type I error very low by rarely rejecting the null hypothesis. Unfortunately this tends to make you keep the null hypothesis when it is wrong. If you reject the null hypothesis often, you will often do so incorrectly, but your probability of Type II error will be decreased. Usually, one decides to limit the probability of Type I error to particular value, or small range of values, and let the Type II error take care of itself. (The reason is the following. Imagine you are conducting a criminal trial. The null hypothesis, at least in our justice system, is that the defendant is innocent–this is our default position–what we will assume unless convincing evidence of guilt is presented. Our justice system considers it far worse to convict an innocent person–a type I error–than to let a guilty one go free-a type II error. This analogy is useful throughout hypothesis testing.) One can reduce the probability of Type II error for a fixed value of α by increasing the sample size. The larger n is, the more sure you are that you know the population mean, and the less likely you are to make an error. The most common significance level is $\alpha = .05$. That is, you have only a 5% chance of rejecting H_0 when it is true. There are disadvantages to rigidly selecting the value of α in advance, as we shall see later.

- **Step IV: Compute the p-value:** The **p-value** is the probability under the null hypothesis, of seeing a value as or more extreme than the value you actually observed. For instance, suppose in the coin example, you observed 95 heads. The values 96, 97, 98, 99, 100, 0, 1, 2, 3, 4, and 5 heads are all as far (or further) from the expected number, 50, as the observed value 95. The p-value is the probability of observing 0-5 or 95-100 heads **if the coin is fair.** If you did the computation, you would see that this is a very small probability indeed. The p-value tells you how likely or unlikely what you observed is if the null hypothesis is true. Thus a very small p-value suggests that what you saw is unlikely, and hence that the null hypothesis is false. A high p-value suggests that what you saw is likely, and therefore consistent with the null hypothesis. (A high p-value does NOT prove that the null hypothesis is TRUE. One example can never prove something is true.)

- **Step V: Decide whether to accept or reject the null hypothesis based on the p-value.** If you specified a significance level α, all you have to do is compare the p-value to α. If the p-value is smaller than α you reject H_0. If it is greater than α you fail to reject H_0. What is a small p-value? The usual rule of thumb is that a p-value of .05 or smaller is small enough to reject the null hypothesis. This is why $\alpha = .05$ is commonly used as the significance level for a test. However, context should

really determine what is a small p-value. If someone's life depends on your not rejecting H_0 when H_0 is true, you want a miniscule p-value. If it's not too important, a p-value around .1-.15 may be fine. I personally prefer not to specify α in advance. My personal rough rule of thumb is the following:

p-value $< .1$: No evidence to reject H_0.
p-value between .05 and .1: Weak evidence to reject H_0.
p-value between .01 and .05: Moderate evidence to reject H_0.
p-value $< .01$: Strong evidence to reject H_0.

One of the great virtues of the p-value is that it lets you perform an hypothesis test for all values of α at once! The p-value is also what is given to you by computer programs, so you MUST be able to interpret it.

There are several standard pairs of null and alternative hypotheses for a single mean or proportion. I will list them briefly here. There are others for more than one mean or proportion, but we will not worry about them in this class.
(1) $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$
(2) $H_0 : \mu \geq \mu_0$ vs $H_A : \mu < \mu_0$
(3) $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$
(4) $H_0 : p = p_0$ vs $H_A : p \neq p_0$
(5) $H_0 : p \geq p_0$ vs $H_A : p < p_0$
(6) $H_0 : p \leq p_0$ vs $H_A : p > p_0$

Examples

Example 1: A company wants to decide whether a particular training program speeds up the performance of the employees on their assembly line. They select a sample of n=25 employees, record the time it takes them to do a task, have them take the training program, and then time them at the task again. For each employee, they compute the reduction in time taken to do the task. The mean reduction for the sample is $\bar{X} = 2.5$ minutes and the sample standard deviation is $s = 9$ minutes. That is, on average, the employees do the task 2.5 minutes faster after training. Test the hypothesis that the training improves efficiency at a level $\alpha = .05$.

Solution:
Steps I and II: The company wants to know if the program helps its employees. The alternative is that it does nothing, or even slows them down. Thus, this is a **one-sided** hypothesis test. The mean reduction time corresponding to the test being useless is 0 minutes. Thus our null and alternative hypotheses are:

$H_0 : \mu \leq 0$
$H_A : \mu > 0$

Step III: The significance level was specified in the problem as $\alpha = .05$.

Step IV: We will be inclined to doubt the null hypothesis that the test has a negative effect if we get high positive reduction times, that is if $\bar{X}$ is large and positive. Note that large negative values only reinforce the null hypothesis. That is why this is called a one-sided test. The p-value is defined as the probability, under H_0 of seeing something more extreme than what you observed, that is something that favors H_A even more than what you observed. This is a one-sided test. Only high positive values favor H_A, so we compute the probability under H_0 of getting a value of $\bar{X}$ **greater than** 2.5:

$$p - value = P(\bar{X} \geq 2.5) = P(\frac{\bar{X} - \mu}{s/\sqrt{n}} \geq \frac{2.5 - 0}{9/\sqrt{25}}) = P(t_{24} \geq 1.389) \approx .08$$

This probability has to be interpolated from the t-table or computed exactly on MINITAB.

We fail to reject the null hypothesis at significance level $\alpha = .05$ because the p-value is greater than our specified significance level $\alpha = .05$ Thus the data is consistent with training program not being helpful at level $\alpha = .05$. However, this is a fairly close call. The p-value is .08. Thus if we had been given a significance level of $\alpha = .1$ we would have rejected the null hypothesis and concluded that the training program is useful. Since this is not a life and death situation, we are probably willing to risk rejecting the null hypothesis fairly frequently when it is right. If I were the company managers, based on the p-value, I would be inclined to keep trying the training program. This is why p-values are more useful than significance levels. If you know the p-value, you know whether to reject or accept the null hypothesis at ANY significance level.

Example 2: Last year, a city measured automobile pollution at a variety of locations and found a mean pollution measurement of 132. This year, they measured at a random sample of n=8 locations and found an average pollution score of $\bar{X} = 120$. The sample standard deviation was found to be s=10. City officials want to know whether their pollution levels are better this year than last year. Test at significance level $\alpha = .05$.

Solution: The null hypothesis is that nothing has happened, that is the pollution level remains unchanged, or has even gotten worse. The alternative–the thing the city officials want to prove– is that pollution levels this year are better. Mathematically this can be stated as

H_0: $\mu \geq 132$
H_A: $\mu < 132$

We are supposed to perform the test at a significance level of $\alpha = .05$. Since we are looking at whether or not pollution levels have improved, a very low value of $\bar{X}$ would make us doubt the null hypothesis.

The p-value is the probability, under H_0 of seeing a value of $\bar{X}$ even more extreme, or more favoring H_A than our observed value of 120. That means we want

$$p - value = P(\bar{X} \leq 120) = P(\frac{\bar{X} - \mu}{s/\sqrt{n}} \leq \frac{120 - 132}{10/\sqrt{8}}) = P(t_7 \leq -3.394) = P(t_7 \geq 3.394) \approx .006$$

At significance level $\alpha = .05$ we reject the null hypothesis that the pollution level is 132 or more. This is because the p-value we computed is less than .05. In fact we would reject the null hypothesis at $\alpha = .025$ or $\alpha = .02$ or any value greater than .006. We would fail to reject the null hypothesis at level $\alpha = .005$ however. This gives us a pretty good idea of how certain we are that the pollution level in this city has changed.

Example 3: In the past, and average household has purchased 5.5 quarts of laundry detergent per year. A government board which monitors consumption of various products wants to know if the amount of laundry detergent used by Americans has changed in the last 20 years. The special value of mu in this case is 5.5. The board wishes to test whether $mu = 5.5$–that is consumption levels have remained unchanged–or whether $\mu \neq 5.5$–i.e. consumption levels have changed. It takes a sample of n=30 households and finds that their average detergent usage over the past year has been $\bar{X} = 5.10$ and the standard deviation of their detergent use is $s = .9$. Perform the appropriate test at level $\alpha = .01$.

Solution: This is a two sided test. Our null hypothesis is that nothing has changed, i.e. $H_0 : \mu = 5.5$. Our alternative is that detergent consumption has changed, i.e. $H_A : \mu \neq 5.5$. As a result either very large OR very small values of $\bar{X}$ will make us doubt the null hypothesis.

Computing the p-value for a two sided test is a bit trickier than for a one-sided test. Recall that the p-value is the probability of seeing something more extreme (more favoring the alternative hypothesis) than what you saw. For a two-sided test, you can have values more extreme than what you observed on EITHER

THE POSITIVE OR NEGATIVE SIDE. Thus the p-value for this problem is $P(\bar{X} \leq 5.1) + P(\bar{X} \geq 5.9)$. Note that by symmetry these two probabilities are the same, so you can just look up the first one in the t-table and DOUBLE the result. The p-value for a two-sided test is usually twice as big as the p-value for a corresponding one-sided test. Here we compute

$$p - value = 2P(\bar{X} \leq 5.1) = 2P(\frac{\bar{X} - \mu}{s/\sqrt{n}} \leq \frac{5.1 - 5.5}{.9/\sqrt{30}}) = 2P(t \leq -2.434) = 2(.010662) = .021322$$

We fail to reject the null hypothesis at level $\alpha = .01$.

Example 4: A manufacturer takes a sample of size n=400 from the company's production line, and finds that 12 of the parts are defective. Company protocols call for producing no more than 2% defective items. Test whether the company is meeting this standard.

Solution: This problem deals with sample proportions. There is some true proportion, p, of items which are defective. The observed proportion in our sample of 400 is $\hat{p} = .03$.

Our null hypothesis is that the production process is fine. That is, 2% or fewer defectives are being produced. The alternative is that too many defectives are being produced. (Why? Because if something is going wrong it is important that we find it out and fix it!) Mathematically this can be stated as

H_0: $p \leq .02$
H_A: $p > .02$

We were not given a significance level in the problem statement. We will be suspicious of the null hypothesis if the proportion of defectives in our sample is very large, that is $\hat{p}$ is large. Note that small values of $\hat{p}$ favor the null hypothesis. Thus, this is a one-sided test. The p-value is

$$p - value = P(\hat{p} \geq .03) = P(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{.03 - .02}{\sqrt{(.02)(.98)/400}}) = P(Z \geq 1.429) = .0764$$

Once again, this is a fairly small p-value. We would reject the null hypothesis at a significance level of $\alpha = .1$. Thus we may be worried that the production process is not up to standard. Another way of saying this is that IF the production process is producing 2% or fewer defectives, there is only about a 7.6% chance of seeing as many defectives as we observed in our sample of 400. This is a pretty small chance!

Example 5: A company is marketing a new product with two different package designs. They want to know if one design is preferred over the other. They ask n=500 people and find that 200 of those surveyed (or 40%) prefer package 1. Is there a difference between the two packages? Use $\alpha = .05$.

Solution: If there were no difference, then the proportion preferring package 1 should be p=.5. This is our null hypothesis. If there is a difference, we would have $p \neq .5$. This is our alternative, because it is more important. If one package is preferred then maybe the company should use that packaging more. This is a two-sided test involving proportions. Our p-value is twice the one-sided p-value or

$$p - value = 2P(\hat{p} \leq .4) = 2P(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq \frac{.4 - .5}{\sqrt{(.5)(.5)/500}}) = 2P(Z \leq -4.47) = .000008$$

I got the exact p-value off the computer. It is extremely small, so we reject the null hypothesis. There is definitely a difference between the two package designs!

Summary

To summarize, here are the possible tests you are responsible for and formulas for computing the corresponding p-values. I let $\bar{X}_{obs}$ and $\hat{p}_{obs}$ stand for the observed sample mean or proportion. The instructions for

performing these tests on the computer are given in MINITAB handout number 2. Note that I have given the formulas using the t distribution when doing tests for the mean. This is because you almost always have s rather than σ. If you have σ instead of s you can use Z where I have t. For proportions we always use Z because under the assumptions of the null hypothesis we actually know the true p and hence the true standard deviation. Note that this is different from confidence intervals for proportions where we do NOT assume we know p and therefore use $\hat{p}$ in the standard deviation formula.

H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$: Reject if $\bar{X}_{obs}$ is large. The p-value is

$$P(\bar{X} \geq \bar{X}_{obs}) = P(t \geq \frac{\bar{X}_{obs} - \mu_0}{s/\sqrt{n}})$$

H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$: Reject if $\bar{X}_{obs}$ is small. The p-value is

$$P(\bar{X} \leq \bar{X}_{obs}) = P(t \leq \frac{\bar{X}_{obs} - \mu_0}{s/\sqrt{n}})$$

H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$: Reject if $\bar{X}_{obs}$ is too large OR small. The p-value is

$$2P(t \geq |\frac{\bar{X}_{obs} - \mu_0}{s/\sqrt{n}}|)$$

H_0: $p \leq p_0$ versus H_A: $p > p_0$: Reject if $\hat{p}_{obs}$ is large. The p-value is

$$P(\hat{p} \geq \hat{p}_{obs}) = P(Z \geq \frac{\hat{p}_{obs} - p_0}{\sqrt{p(1-p)/n}})$$

H_0: $p \geq p_0$ versus H_A: $p < p_0$: Reject if $\hat{p}_{obs}$ is small. The p-value is

$$P(\hat{p} \leq \hat{p}_{obs}) = P(Z \leq \frac{\hat{p}_{obs} - p_0}{\sqrt{p(1-p)/n}})$$

H_0: $p = p_0$ versus H_A: $p \neq p_0$: Reject if $\hat{p}_{obs}$ is too large OR small. The p-value is

$$2P(Z \geq |\frac{\hat{p} - p_0}{\sqrt{p(1-p)/n}}|)$$